electronics

MDPI

Article

Think Before You Classify: The Rise of Reasoning Large Language Models for Consumer Complaint Detection and Classification

Konstantinos I. Roumeliotis [1,2,*], Nikolaos D. Tselikas [1] and Dimitrios K. Nasiopoulos [3]

[1] Department of Informatics and Telecommunications, University of the Peloponnese, 22131 Tripoli, Greece; ntsel@uop.gr

[2] Department of Digital Systems, University of the Peloponnese, 23100 Sparta, Greece

[3] Department of Agribusiness and Supply Chain Management, School of Applied Economics and Social Sciences, Agricultural University of Athens, 11855 Athens, Greece; dimnas@aua.gr

* Correspondence: k.roumeliotis@uop.gr

Abstract: Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing (NLP) tasks, but their effectiveness in real-world consumer complaint classification without fine-tuning remains uncertain. Zero-shot classification offers a promising solution by enabling models to categorize consumer complaints without prior exposure to labeled training data, making it valuable for handling emerging issues and dynamic complaint categories in finance. However, this task is particularly challenging, as financial complaint categories often overlap, requiring a deep understanding of nuanced language. In this study, we evaluate the zero-shot classification performance of leading LLMs and reasoning models, totaling 14 models. Specifically, we assess DeepSeek-V3, Gemini-2.0-Flash, Gemini-1.5-Pro, Anthropic's Claude 3.5 and 3.7 Sonnet, Claude 3.5 Haiku, and OpenAI's GPT-4o, GPT-4.5, and GPT-4o Mini, alongside reasoning models such as DeepSeek-R1, o1, and o3. Unlike traditional LLMs, reasoning models are specifically trained with reinforcement learning to exhibit advanced inferential capabilities, structured decision-making, and complex reasoning, making their application to text classification a groundbreaking advancement. The models were tasked with classifying consumer complaints submitted to the Consumer Financial Protection Bureau (CFPB) into five predefined financial classes based solely on complaint text. Performance was measured using accuracy, precision, recall, F1-score, and heatmaps to identify classification patterns. The findings highlight the strengths and limitations of both standard LLMs and reasoning models in financial text processing, providing valuable insights into their practical applications. By integrating reasoning models into classification workflows, organizations may enhance complaint resolution automation and improve customer service efficiency, marking a significant step forward in AI-driven financial text analysis.

Keywords: reasoning models; reasoning model classification; large language models; consumer complaints; consumer complaint classification; zero-shot classification; deepseek model; gpt model; claude model; gemini model

Academic Editor: Jian Sun

Received: 11 February 2025
Revised: 3 March 2025
Accepted: 5 March 2025
Published: 7 March 2025

Citation: Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Think Before You Classify: The Rise of Reasoning Large Language Models for Consumer Complaint Detection and Classification. *Electronics* **2025**, *14*, 1070. https://doi.org/10.3390/electronics14061070

1. Introduction

Understanding consumer complaints is crucial for businesses, regulators, and financial institutions. Effective complaint management not only improves customer satisfaction but also helps organizations identify systemic issues, ensure compliance, and enhance decision-making [1]. The Consumer Financial Protection Bureau (CFPB) serves as a key regulatory body in the U.S., collecting consumer complaints related to financial products and services [2]. These complaints, often submitted as free-text narratives, contain valuable insights but require accurate classification to be processed efficiently. The challenge lies in categorizing these complaints into well-defined financial classes, as they often involve overlapping categories and complex financial terminology [3].

The classification of consumer complaints is a demanding natural language processing (NLP) task due to the unstructured nature of financial narratives and the nuanced distinctions between complaint types [4]. Traditional approaches have relied on machine learning (ML), deep learning (DL), and NLP models trained on annotated financial datasets to automate complaint classification, improving efficiency over manual tagging [3,5]. However, as large language models (LLMs) continue to evolve, the prospect of performing this classification in a zero-shot setting—without additional task-specific training—has become increasingly viable [6].

This study explores the effectiveness of 14 state-of-the-art LLMs in zero-shot consumer complaint classification. We evaluate a diverse set of models, including DeepSeek-V3 [7], OpenAI's GPT-4o, GPT-4.5, and GPT-4o mini [8], and Anthropic's Claude 3.5 Sonnet, Claude 3.7 Sonnet, and Claude 3.5 Haiku [9], as well as Google's Gemini-2.0-flash, Gemini-2.0-flash-lite, Gemini-1.5-pro, and Gemini-1.5-flash [10]. These models are among the most advanced in AI and are known for their strong language comprehension and text classification capabilities [11].

Beyond standard LLMs, our study introduces a novel approach by evaluating reasoning models—DeepSeek-R1, o1, and o3—in the context of classification tasks. Unlike traditional classification-focused LLMs, reasoning models are specifically trained with reinforcement learning to perform complex reasoning, problem-solving, coding, scientific analysis, and multi-step planning for agentic workflows [12]. These models exhibit advanced problem-solving abilities, logical inference, and structured decision-making. While such capabilities have been explored in areas such as mathematical reasoning, commonsense understanding, and complex multi-step problem-solving, their application to classification tasks remains largely uncharted.

Integrating reasoning models into classification tasks represents a paradigm shift in NLP. Instead of relying solely on pattern recognition and statistical learning, reasoning models can analyze textual nuances, infer implicit relationships, and resolve ambiguous complaint narratives with a structured, step-by-step approach. This could be particularly valuable in financial complaint classification, where overlapping categories and contextual dependencies often lead to misclassification. By leveraging reasoning-based inference, these models may enhance classification accuracy and explainability, offering a more robust alternative to traditional LLM-based classifiers.

To assess the zero-shot classification performance of these models, we evaluate how each one categorizes consumer complaints based solely on the complaint text. We employ key performance metrics, including accuracy, precision, recall, and F1-score, to compare model effectiveness. Additionally, we generate heatmaps to visualize classification patterns and identify strengths and weaknesses across different models.

Our primary research question is: how well do top-tier LLMs and reasoning models perform in zero-shot consumer complaint classification, and which model achieves the highest accuracy in categorizing financial disputes? By addressing this question, we aim to provide insights into the practical application of state-of-the-art LLMs and reasoning models in financial text classification, highlighting their potential to redefine automated complaint processing.

2. Literature Review

Consumer complaints are prevalent across various sectors, including financial services, e-commerce, transportation, and environmental concerns. Efficiently managing these complaints is crucial for businesses to enhance user satisfaction, build trust, and retain customers [13]. As digital platforms become the primary medium for consumer interactions, organizations must process vast complaint volumes effectively. Manual handling is often impractical due to sheer volume, making automated classification and prioritization essential. Complaint classification categorizes grievances into structured groups, streamlining customer service operations and aiding in sentiment analysis to improve service quality [14].

2.1. NLP-Based Complaint Analysis

Vairetti et al. (2024) [3] propose a DL and multi-criteria decision-making (MCDM) framework for complaint prioritization. Their study emphasizes efficient categorization to improve customer satisfaction (CSAT) and reduce churn rates. Using a BERT-based model, BETO, they achieve a 92.1% accuracy rate, showcasing the effectiveness of modern NLP models in complaint classification.

Kumar et al. (2024) [15] examine consumer complaints in multilingual e-commerce settings, integrating sentiment, emotion, and severity analysis. Their hierarchical attention-based DL model processes complaints at word and sentence levels, outperforming benchmark models and emphasizing the importance of emotional and severity considerations in complaint handling.

Seok et al. (2024) [16] introduce a DL-based customer complaint monitoring system using explainable AI (XAI) techniques. Their approach integrates BERT-based models to extract service-related features from online reviews and analyze sentiment. Their study employs a staged p-chart for continuous monitoring of complaints in seasonal industries, addressing limitations in traditional NLP techniques.

Khadija et al. (2024) [17] apply latent Dirichlet allocation (LDA) combined with BERT embeddings to analyze Indonesian customer complaints. Their findings suggest that BERT-based models enhance LDA's topic coherence and silhouette scores, making them effective for extracting meaningful topics from short-text complaints, thus aiding businesses in structured complaint categorization.

Sharma et al. (2024) [4] introduce a multimodal NLP feedback system that enables silent feedback collection in shopping malls. Their model processes audio feedback to extract customer sentiments, reducing communication barriers and improving feedback quality.

2.2. Generative AI (GAI) in Complaint Handling

Juipa et al. (2024) [18] introduce a sentiment analysis-based chatbot for telecommunications complaint management. Their study demonstrates that incorporating GPT-3.5 for sentiment analysis improves complaint resolution efficiency and customer satisfaction. The chatbot achieves an 86% satisfaction rate, exceeding industry standards and supporting AI-driven chatbots in enhancing complaint-handling processes.

Das et al. (2024) [19] focus on financial complaints, proposing an explainable AI approach to differentiate between negative reviews and formal complaints. Their dyadic attention-based model facilitates sentiment detection, emotion recognition, and severity classification, offering a comprehensive understanding of customer dissatisfaction in financial services.

Jia et al. (2024) [20] explore generative AI (GAI) in customer complaint resolution, comparing AI-generated responses with human-authored ones. Their research highlights GAI's potential for automating complaint responses while maintaining empathy and coherence, emphasizing the need for balancing AI integration with human oversight to ensure authentic interactions. Correa et al. (2024) [21] extend the role of generative AI in consumer complaint handling, integrating classification, summarization, and response generation, achieving an 88% classification accuracy and demonstrating AI's potential in customer service operations.

2.3. Machine Learning (ML) and Deep Learning (DL) Models for Complaint Management

Roy et al. (2024) [5] explore complaint classification in the railway sector, highlighting the role of social media in customer feedback. Their system employs ML models such as random forest and support vector machines (SVMs) to classify tweets based on urgency, ensuring timely responses to critical complaints. Sentiment analysis plays a key role in identifying negative tweets that require immediate attention.

Jondhale et al. (2024) [22] developed an ML-based pipeline for predicting disputed financial complaints. Utilizing PySpark ML and feature engineering, their system enhances complaint classification accuracy and supports proactive dispute resolution.

Djahongir Ismailbekovich (2024) [23] investigates AI-driven chatbots for consumer complaints, noting their efficiency benefits. However, challenges such as bias, lack of nuance interpretation, and regulatory compliance necessitate a hybrid approach that combines AI automation with human oversight.

Song et al. (2024) [24] propose a textual analysis framework integrating guided LDA and sentiment polarity for quantifying service failure risks. By incorporating CRITIC and TOPSIS methodologies, their approach enhances traditional failure mode and effects analysis (FMEA), offering a data-driven risk assessment method. Applied to the hotel industry, their findings improve accuracy in identifying and prioritizing service failure risks.

Wang et al. (2024) [25] analyze consumer complaint behavior in live-streaming e-commerce using a two-staged SEM-ANN approach. Their study identifies key complaint factors, including consumer confusion, emotional venting, and altruistic motives. The research highlights the impact of group complaints on consumer behavior and suggests strategies for e-commerce platforms to mitigate negative experiences.

2.4. Broader Implications of Consumer Complaints

Zhou et al. (2024) [26] analyze environmental complaints in China, illustrating how consumer feedback influences public policy and environmental quality. Their study underscores the role of complaint reporting in pollution control efforts, broadening the scope of complaint management applications.

Lee et al. (2024) [27] developed the "3D" model for temperament-centered complaints. Their e-Customer Complaint Handling (e-CCH) system collects, processes, and classifies complaints based on consumer temperaments, providing personalized solutions. Their open-sourced dataset offers valuable insights for industrial service management and complaint-handling research.

Overall, the literature underscores the growing reliance on AI, ML, and NLP for consumer complaint classification and sentiment analysis. These advancements have significantly improved accuracy, efficiency, and scalability in handling consumer grievances across industries.

3. Materials and Methods

This section outlines the systematic approach employed to explore and evaluate the capabilities of 14 different models, including DeepSeek-V3, GPT-4o, Gemini-2.0-Flash, and Claude-3.5-Sonnet, along with their reasoning models in the challenging task of zero-shot consumer complaint classification. Given the complexity of financial complaints—where categories often overlap and subtle linguistic nuances influence classification—this study follows a structured methodology to ensure a fair and comprehensive assessment of each model's performance.

We begin by describing the dataset used in our analysis, which consists of consumer complaints submitted to the CFPB [2]. Since these complaints are written in free-text form, preprocessing is a crucial step to standardize and prepare the data for classification. In the preprocessing steps, we detail the cleaning procedures applied to the complaint narratives to improve input consistency across all models.

Next, we discuss prompt engineering, a critical component of zero-shot classification. As these models have not been fine-tuned for this specific task, the prompt structure plays a significant role in guiding the models toward accurate classifications. We describe the design and refinement of prompts tailored to extract the best possible performance from each model.

Finally, we present our model evaluation strategy, defining the metrics used to assess classification accuracy. The performance of LLMs and their reasoning models is measured using standard classification metrics, including accuracy, precision, recall, and F1-score. Additionally, heatmaps are employed to analyze misclassification patterns and compare trends across models.

By implementing a rigorous methodology, this study aims to provide a data-driven comparison of LLMs and reasoning models in real-world consumer complaint classification, offering valuable insights into their effectiveness for financial text processing.

3.1. Description of the Dataset

The dataset used in this study is the Consumer Complaints Dataset for NLP, curated by Shashwat Tiwari and hosted on Kaggle [28]. This dataset is derived from consumer complaints originally sourced from the CFPB website [2]. It contains consumer-submitted complaints about financial products and services, making it a highly relevant resource for evaluating LLMs in the demanding task of financial text classification.

The dataset includes one year's worth of consumer complaints, covering the period from March 2020 to March 2021. Additionally, the dataset's creator supplemented the complaints by utilizing the CFPB's API to fetch up-to-the-minute submissions, ensuring a mix of historical and more recent complaints. Each complaint is associated with one of nine original financial product categories, but due to similarities between certain classes and class imbalances, the dataset has been consolidated into five broader financial categories: Credit Reporting, Debt Collection, Mortgages and Loans (includes car loans, payday loans, student loans, etc.), Credit Cards, and Retail Banking (includes checking/savings accounts, money transfers, Venmo, etc.).

The dataset consists of approximately 162,400 consumer complaints, each containing a free-text narrative describing the consumer's financial issue. The text length varies significantly, with an average length of 588.49 characters and a maximum length of 20,596 characters, posing a challenge for classification models.

Additionally, the dataset is highly imbalanced, with the following distribution:

- Credit Reporting: 56.14%
- Debt Collection: 14.25%
- Mortgages and Loans: 11.69%
- Credit Cards: 9.58%
- Retail Banking: 8.33%

Given its 9.41 usability score on Kaggle, this dataset is widely regarded as a valuable benchmark for consumer complaint classification using NLP. By leveraging this dataset, we aim to evaluate the zero-shot classification capabilities of LLMs, assessing their ability to process real-world financial complaints efficiently and accurately.

3.2. Preprocessing Steps

To ensure a high-quality dataset for evaluating both the LLMs and their reasoning models in zero-shot consumer complaint classification, several preprocessing steps were applied. These steps aimed to improve data consistency, address class imbalance, and optimize the dataset for model evaluation. Given the complexity and overlapping nature of financial complaints, preprocessing was crucial in ensuring fair and meaningful model comparisons.

3.2.1. Removing Excessively Long Narratives

The dataset includes consumer complaints of varying lengths, with some narratives reaching 20,596 characters. To enhance computational efficiency and prevent outliers from skewing classification results, complaints exceeding 500 characters were removed. This step ensured that all models, including DeepSeek-V3, GPT-4o, Gemini-2.0-Flash, and Claude-3.5-Sonnet, processed narratives of a more typical length, aligning with real-world classification constraints.

3.2.2. Standardizing Narrative Text

To enhance textual consistency, the narrative column underwent standardization, including:

- Converting all text to lowercase
- Removing special characters, excessive whitespace, and inconsistent formatting
- Normalizing the text's structure to ensure cleaner inputs
- This standardization prevented inconsistencies that could impact classification accuracy and helped models to better understand the content of complaints.

3.2.3. Removing Entries with Empty Fields

Some complaints were missing key information in critical columns such as "product" and "narrative". These incomplete entries were removed to maintain data integrity and ensure that each record contained a valid complaint description and category. This step eliminated potential biases arising from incomplete data affecting model predictions.

3.2.4. Creating a Balanced Subset Using Undersampling

The original dataset was highly imbalanced, with credit reporting complaints making up over 56% of submissions, while retail banking complaints accounted for only 8.33%. To create a balanced evaluation set, an undersampling technique was applied, selecting a stratified subset of 1000 records. The resulting dataset ensured equal representation across all five complaint categories, with:

- 200 complaints per category
- 20.00% representation per category

This balanced subset allowed for an unbiased assessment of models in zero-shot classification, ensuring that no single category disproportionately influenced model performance.

Through these preprocessing steps, the dataset was refined to provide clean, standardized, and balanced inputs, supporting a robust comparison of LLMs and reasoning models in financial complaint classification.

3.3. Prompt Engineering

Crafting effective prompts is essential for optimizing the performance of LLMs, particularly in zero-shot classification tasks where the model has not been specifically fine-tuned. This technique involves structuring input queries in a way that guides the model to generate precise and relevant responses. The effectiveness of a prompt largely determines how well the model understands the task and produces accurate results, making prompt design a crucial skill for leveraging LLMs efficiently.

In this research, prompt engineering played a key role in enabling models to classify consumer complaints into predefined categories based only on the complaint text, without prior exposure to those classifications. Since LLMs are inherently flexible, developing effective prompts required an iterative process, where refinements were made to enhance clarity, reduce ambiguity, and improve classification accuracy. Elements such as the specificity of instructions, data presentation format, and overall structure were carefully adjusted to optimize real-world applications [29].

To establish an effective prompt, we conducted a trial-and-error process using the GPT model's chat interface. Through multiple refinements, we devised a structured prompt that improved interpretability. Given that structured input generally enhances LLM performance, we initially experimented with XML formatting to organize the information. While this approach made the data more accessible to the model, it also increased the token count, making API interactions more resource-intensive and expensive.

To address these limitations, we leveraged Anthropic's console prompt generator—a tool designed to help users craft effective prompts by providing pre-built templates based on best practices [30]. Our experimentation revealed that JSON formatting was a more efficient alternative to XML, as it streamlined the prompt structure and reduced token usage, making the process more cost-effective.

However, ensuring consistency in prompt structure across different LLMs presented challenges, as models often interpret formatting and instruction nuances differently. Aligning prompt design to maintain cross-model performance required balancing universality and efficiency while considering each model's response tendencies.

The final prompt was carefully refined to strike a balance between clarity and efficiency, ensuring it was universally applicable across different LLMs while minimizing computational overhead. Listing 1 illustrates an example of the optimized prompt. This structured approach allowed us to enhance model performance while effectively managing API constraints.

Listing 1. Model-agnostic prompt.

```
conversation.append({
'role': 'user',
'content':
        'You are an AI assistant specializing in consumer complaint classification.'
        'Your task is to analyze a consumer complaint and classify it into the most'
        'appropriate category from the predefined list:'
        '["retail_banking", "credit_reporting", "credit_card", "mortgages_and_loans",
"debt_collection"]'
        'Provide your final classification in the following JSON format without
explanations:'
        '{"product": "chosen_category_name"}.\nComplaint: '
        '...'
})
```

3.4. Model Predictions

For this study, we utilized the GPT-4o, GPT-4.5-preview-2025-02-27, o1-2024-12-17, GPT-4o-mini, o3-mini-2025-01-31, Claude-3-5-Sonnet-20241022, Claude-3-7-Sonnet-20250219, Claude-3-5-Haiku-20241022, DeepSeek-Chat, DeepSeek-Reasoner, Gemini-2.0-Flash, Gemini-2.0-Flash-Lite, Gemini-1.5-Pro, and Gemini-1.5-Flash models to perform zero-shot classification of consumer complaints. The goal was to assess how well each model could categorize financial disputes without fine-tuning, relying solely on their pre-trained knowledge.

To ensure a structured and efficient evaluation process, we developed a reusable Python class that systematically handled model interactions while separating the core logic from the dataset. This class performed the following key operations:

- Iterating through the dataset—Each complaint was processed row by row.
- Constructing the prompt—The complaint text was dynamically combined with the predefined complaint categories to generate a structured prompt for classification.
- Managing API communications—The appropriate API was called for each respective model, ensuring seamless interaction.
- Handling responses—The models were instructed to return their predictions in JSON format. However, DeepSeek-V3 frequently included extra text outside the expected JSON structure, requiring additional processing to extract and clean the valid classification output. A separate method was implemented to search for and capture the valid JSON response, ensuring uniformity across all model outputs.
- Storing results—All predictions were saved within the same dataset file, making it easier to analyze classification performance across different models.

Additionally, to measure the computational efficiency of each model, we recorded the time taken for each prediction and stored this information in the same CSV file. This allowed us to compare not only accuracy but also latency, which is a crucial factor in real-world applications.

In line with our commitment to open science and unbiased research, we have made all code available as an open-source project on GitHub, licensed under MIT, allowing researchers and developers to replicate and extend our work [31].

3.5. Model Evaluation Strategy

To assess the effectiveness of DeepSeek, GPT, Gemini, and Claude models, along with their reasoning models, in zero-shot consumer complaint classification, a comprehensive evaluation strategy was implemented. The primary goal was to measure each model's classification accuracy across the five predefined categories and analyze performance trends.

3.5.1. Accuracy-Based Comparison

A straightforward metric for evaluation was exact match accuracy, which counts how many predictions exactly matched the true complaint category. This method provides a quick and direct comparison of classification performance across all models.

Beyond simple accuracy, four key classification metrics were computed for each model:

- Accuracy—The proportion of correctly classified complaints.
- Precision (weighted)—How often a model's predicted category was correct, considering class imbalances.
- Recall (weighted)—How well the model identified complaints belonging to each category.
- F1-score (weighted)—The harmonic mean of precision and recall, balancing both metrics.

Each model's evaluation results were stored in a dedicated CSV file, allowing easy comparison and further analysis.

3.5.2. Confusion Matrix and Heatmap Analysis

To visualize classification patterns, a heatmap was generated based on a confusion matrix. This helped identify:

- Common misclassifications—Whether certain categories were frequently confused with others.
- Model biases—If a model tended to favor certain categories over others.

These heatmaps provided deeper insights into the strengths and weaknesses of models, highlighting their classification tendencies in financial consumer complaints.

By leveraging these evaluation techniques, we established a quantitative foundation for comparing the zero-shot classification capabilities of these leading LLMs, offering valuable insights into their suitability for real-world financial NLP applications.

4. Results

In Section 3, we detailed the methodology used to prompt the LLMs and reasoning models for this classification task. Here, we evaluate the performance of DeepSeek, GPT, Gemini, and Claude in zero-shot consumer complaint classification by examining their accuracy, precision, recall, F1-score, cost, and inference speed.

This section provides a detailed discussion of the findings, offering insights into each model's capabilities, strengths, and weaknesses. The complete outcomes of our study are summarized in Table 1.

Table 1. Comparison of model performance metrics.

Model	Accuracy	Precision	Recall	F1	Cost (USD)
gpt-4o	0.736	0.7546	0.736	0.7358	0.45
gpt-4.5-preview-2025-02-27	0.774	0.7886	0.774	0.7763	12.32
o1-2024-12-17	0.777	0.7944	0.777	0.7764	24.24
gpt-4o-mini	0.691	0.7026	0.691	0.6933	0.03
o3-mini-2025-01-31	0.752	0.7671	0.752	0.7514	1.72
claude-3.5-sonnet-20241022	0.763	0.7973	0.763	0.7617	0.64
claude-3.7-sonnet-20250219	0.765	0.7872	0.765	0.7626	0.64
claude-3.5-haiku-20241022	0.721	0.7287	0.721	0.7227	0.17
deepseek-chat	0.737	0.752	0.737	0.7368	0.01
deepseek-reasoner	0.738	0.7565	0.738	0.7382	0.26
gemini-2.0-flash	0.758	0.7661	0.758	0.7586	0.04
gemini-2.0-flash-lite	0.742	0.7488	0.742	0.7402	0.01
gemini-1.5-pro	0.753	0.7639	0.753	0.7515	0.23
gemini-1.5-flash	0.707	0.7358	0.707	0.705	0.01

4.1. Classification Performance

4.1.1. Overall Accuracy

Among the LLMs tested, GPT-4.5 and Claude-3.7-Sonnet achieved the highest accuracies of 0.774 and 0.765, respectively, demonstrating their superior ability to classify financial complaints correctly. This suggests that they can better distinguish between overlapping complaint categories without requiring fine-tuning.

Claude-3.5-Sonnet delivered competitive results compared to its successors (0.763), while GPT-4o fell significantly below GPT-4.5 (0.736). The recent Gemini models, such as Gemini-2.0-Flash (0.758) and Gemini-2.0-Flash-Lite (0.742), followed closely behind the leading models. DeepSeek-Chat (0.737) slightly outperformed GPT-4o, indicating its competitiveness in handling financial texts.

However, Claude-3.5-Haiku (0.721), GPT-4o-mini (0.691), Gemini-1.5-Pro (0.753), and Gemini-1.5-Flash (0.707) performed slightly worse, highlighting a trade-off between model size and classification effectiveness.

Nevertheless, reasoning models proved to be the game-changers. The o1 reasoning model outperformed the leading GPT-4.5 LLM (0.777), while the o3-mini model significantly surpassed GPT-4o-mini (0.752), achieving results comparable to high-end models despite its smaller size. In contrast, the DeepSeek-Reasoner model delivered less promising results, performing almost identically to its chat counterpart (0.738) and lagging behind the o1 and o3-mini reasoning models.

4.1.2. Precision, Recall, and F1-Score

Surprisingly, in terms of precision, the Claude-3.5-Sonnet model achieved the highest precision (0.7973), surpassing the o1 reasoning model by 0.29% and its successor, Claude-3.7-Sonnet, by 1.01%. GPT-4.5 follows with 0.7886, while the o3-mini reasoning model surpassed the Gemini-2.0-Flash model in precision by 0.1% and the Gemini-1.5-Pro model by 0.32%.

GPT-4o (0.7546) and DeepSeek-V3 (0.752) exhibited similar precision, positioning them as reliable choices, whereas GPT-4o-mini had the lowest precision (0.7026), indicating a higher likelihood of incorrect classifications.

In terms of recall, which measures how well a model captures all instances of a category, the o1 reasoning model and GPT-4.5 led with 0.777 and 0.774, respectively, followed by Claude-3.7-Sonnet and Claude-3.5-Sonnet with 0.765 and 0.763, reinforcing their robustness.

The F1-score, which balances precision and recall, further confirmed the o1 reasoning model (0.7764) as the best overall performer, with GPT-4.5 (0.7763), Claude-3.7-Sonnet, and Claude-3.5-Sonnet also achieving competitive scores. Once again, GPT-4o-mini had the lowest F1-score (0.6933), reflecting its reduced effectiveness in this classification task.

4.2. Cost Efficiency

All predictions were made using official API versions to ensure a fair comparison. While DeepSeek is open-source and can be run locally, API-based inference was used to standardize the results.

The cost per 1000 classifications varied significantly across models. DeepSeek-V3, Gemini-2.0-Flash-Lite, and Gemini-1.5-Flash were the most cost-efficient at USD 0.01 per 1000 classifications. However, for DeepSeek, this price is part of a limited-time offer valid until 8 February 2025. After this date, costs will increase to USD 0.14 per million input tokens and USD 0.28 per million output tokens, though it will still remain more affordable than GPT-4o and Claude models. GPT-4o-mini and Gemini-2.0-Flash were also highly affordable at USD 0.03 and USD 0.04 per 1000 classifications, respectively.

As expected, reasoning models with a long internal chain of thought before responding were the most expensive. The o1 reasoning model had the highest cost at USD 24.24 per 1000 classifications, followed by the latest GPT-4.5 model at USD 12.32 due to its preview phase. The o3-mini model also employs a long chain of thought, and although much cheaper than the o1 reasoning model, it is still significantly more expensive than the other models. Interestingly, the DeepSeek-Reasoner model, despite being 2500% more expensive than the DeepSeek-Chat model, remains far more affordable than the o1 reasoning model. However, this cost difference appears to come at the expense of accuracy and precision.

It is worth noting that although the DeepSeek-Reasoner model is considerably more expensive than DeepSeek-Chat, their accuracy and precision metrics are nearly identical. This raises concerns about whether DeepSeek-Reasoner is a truly distinct reasoning model or merely a version of DeepSeek-Chat that explicitly describes its thought process before reaching a conclusion.

GPT-4o and Claude 3.5 Haiku offered a balance between cost and performance. GPT-4o was priced at USD 0.45 per 1000 classifications, with a rate of USD 2.50 per million input tokens and USD 10.00 per million output tokens. Claude 3.5 Haiku, at USD 0.17 per 1000 classifications, had a pricing structure of USD 0.80 per million input tokens and USD 4.00 per million output tokens.

Claude 3.7 and Claude 3.5 Sonnet were priced similarly at USD 0.64 per 1000 classifications. Despite their higher cost compared to Gemini and DeepSeek models, the Claude-3.7-Sonnet model delivered the best classification precision, making it a viable choice for scenarios where precision is critical.

For budget-conscious applications, DeepSeek-V3 and Gemini-2.0-Flash-Lite remain the most economical options. However, DeepSeek's significantly slower inference speed must be carefully considered.

4.3. Inference Speed and Latency

Inference time is crucial for real-time classification applications. The mean prediction time and total time for processing 1000 complaints are summarized in Table 2. These results highlight the trade-offs between model speed and classification accuracy, which are essential considerations when selecting an LLM for time-sensitive tasks.

Table 2. Inference Speed Comparison of LLMs.

Model	Mean Prediction Time (Seconds)	Total Time (for 1000 Complaints) (Seconds)
gpt-4o	0.89	889.25
gpt-4.5-preview	1.52	1520.49
o1	4.8	4804.12
gpt-4o-mini	0.86	860.44
o3-mini	4.54	4538
claude-3.5-sonnet	1.8	1797.11
claude-3.7-sonnet	1.12	1123.65
claude-3.5-haiku	1.81	1806.23
deepseek-chat	26.69	26,686.94
deepseek-reasoner	22.56	22,564.02
gemini-2.0-flash	0.68	680.35
gemini-2.0-flash-lite	0.69	691.53
gemini-1.5-pro	0.82	818.9
gemini-1.5-flash	0.72	718.56

Observations:

- Gemini-2.0-Flash and Gemini-2.0-Flash-Lite were the fastest models, with prediction times of 0.68 s and 0.69 s per prediction, respectively. They were followed closely by Gemini-1.5-Flash (0.72 s), GPT-4o-mini (0.86 s), and GPT-4o (0.89 s). These models are ideal for low-latency applications requiring fast classification.
- Claude 3.5 Sonnet and Claude 3.5 Haiku were slower (~1.8 s per prediction), suggesting higher computational demands. However, the latest Claude-3.7-Sonnet model showed improved efficiency, reducing its prediction time to 1.12 s.
- DeepSeek-Chat was the slowest among LLMs, with an average prediction time of 26.69 s per request. This makes it unsuitable for real-time applications despite its low cost. Additionally, DeepSeek's API has faced recent performance issues due to high demand and cyberattacks following its launch on 20 January 2025.

- As expected, the reasoning models o1 and o3-mini, due to their extensive chain of thought, were among the slowest after the DeepSeek models, with mean prediction times of 4.8 s and 4.54 s, respectively.

4.4. Heatmaps: Model Performance Analysis

Heatmaps are powerful visualization tools that aid in analyzing classification tasks by providing an intuitive and comprehensive way to interpret model performance [32]. They visually represent data distributions, feature correlations, and classification results using color gradients, making complex patterns more accessible. In classification tasks, heatmaps are commonly utilized for confusion matrices, feature importance analysis, and activation mapping in DL models, helping researchers and practitioners identify misclassifications, feature relevance, and decision boundaries. In particular, for LLMs, heatmaps can help researchers determine whether a model achieves sufficiently good zero-shot results or requires further fine-tuning. In this study, we employ heatmaps to gain deeper insights into the models' strengths and weaknesses across the five classification categories, as illustrated in Figure 1. During the evaluation phase, the heatmaps are generated using the cross-tabulation method, which creates a contingency table showing the frequency of actual vs. predicted values. This table is then visualized using Seaborn's heatmap function, allowing for an intuitive comparison of classification performance.

Figure 1 displays 14 confusion matrices, each corresponding to a distinct classification model prompt for categorizing consumer complaints into a uniform set of predefined categories. In each subplot, the vertical axis represents the true class while the horizontal axis represents the predicted class. A color scale is used, ranging from light green (indicating relatively few samples) to dark blue (indicating a high number of samples). Darker cells along the main diagonal reflect higher classification accuracy for the respective category, whereas lighter off-diagonal cells indicate elevated misclassification rates. The overall accuracy metric is computed by summing the values along the main diagonal. This visualization enables a clear identification of each model's weaknesses and suggests that targeted fine-tuning with additional labeled data could improve classification performance. For instance, categories such as retail_banking, which exhibit consistently high accuracy across models, may be excluded from further fine-tuning to enhance cost-effectiveness.

An initial observation is derived from the first plot associated with the GPT-4o model, where the model disregarded the explicit classification instructions by predicting the auto_loan category three times, despite this category not being included in the predefined set. This behavior implies the presence of internal biases or a tendency to generalize beyond the given labels. In contrast, its successor, GPT-4.5, does not appear to exhibit the same issue. A similar anomaly is observed in the Gemini-2.0-Flash-Lite plot, where the model incorrectly predicts the student_loans category, which is also not part of the specified prompt.

Among all the models, the dept_collection category appears to be the most challenging, with a high incidence of misclassification—often being erroneously predicted as credit_reporting—suggesting that consumer complaints in this category are inherently difficult to discern. A comparable pattern is noted for the credit_card category, where many models misclassify complaints as retail_banking, likely due to the broader nature of the retail_banking label. Conversely, the credit_reporting and retail_banking categories are among the most accurately classified, as evidenced by the consistently darker cells along the main diagonal in the corresponding confusion matrices.

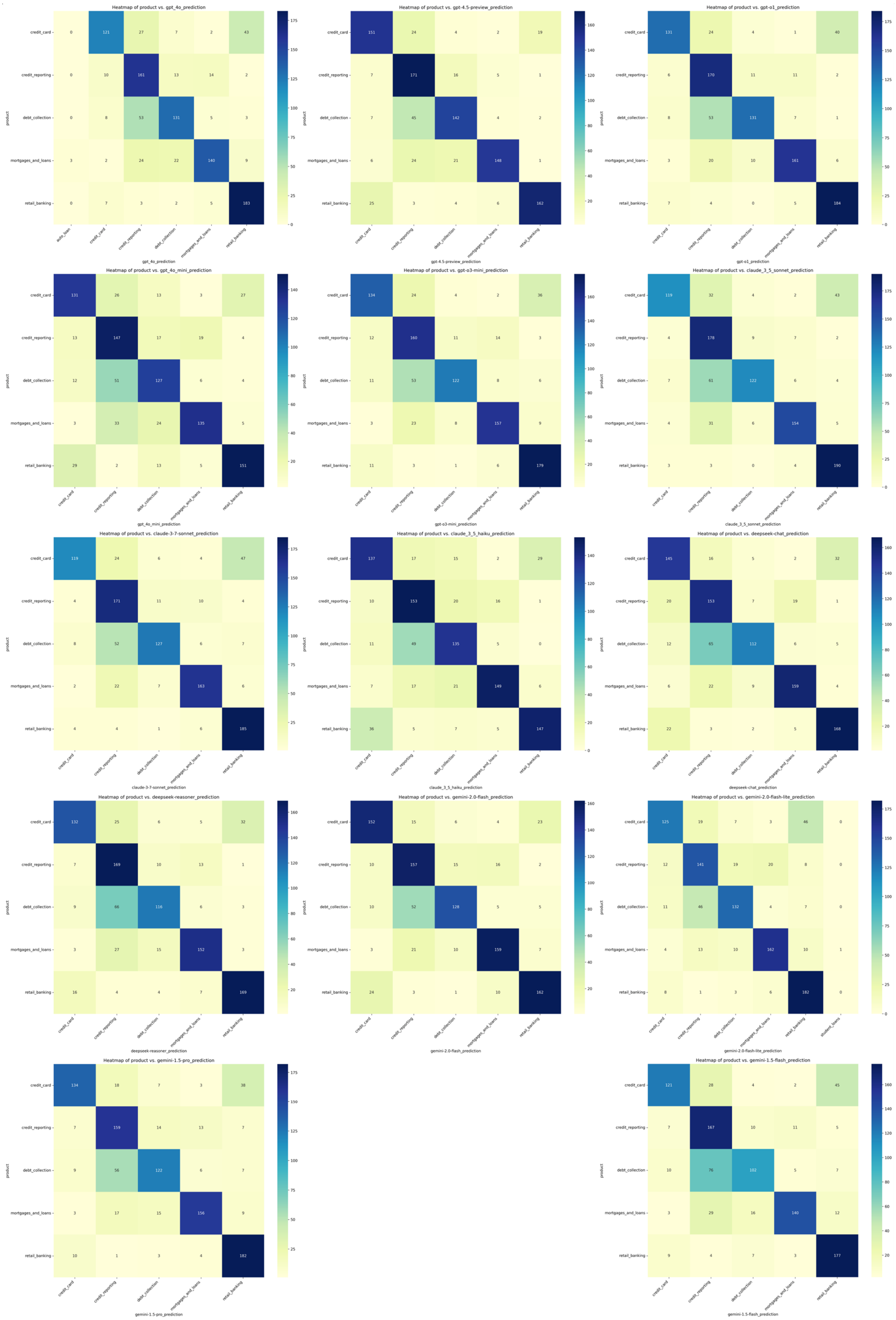

Figure 1. Comparison of Model Predictions with Actual Categories Using Heatmaps.

5. Discussion

5.1. Overview of Classification Performance

The results presented in Section 4 highlight the strong performance of various LLMs in classifying financial complaints. The models successfully categorized closely related financial terms into five distinct categories without task-specific fine-tuning. Given that a random classification approach would yield a 20% accuracy rate (one-fifth probability), the observed accuracy levels exceeding 70% are a remarkable achievement. This performance can be attributed to the extensive pre-training phase of LLMs, which exposes them to a broad spectrum of financial data. However, while 70% accuracy is notable in an experimental setting, it may not be optimal for real-world applications, necessitating further fine-tuning for improved reliability.

5.2. Misclassification Patterns and Biases

Beyond accuracy, the heatmaps provided insights into misclassification tendencies. Notably, GPT-4o misclassified three mortgages_and_loans complaints as auto_loan issues—a category not included in the prompt. The same occurred with the Gemini-2.0-Flash-Lite model, which classified one mortgages_and_loans complaint as student_loans. Interestingly, their successors and counterparts did not exhibit this behavior, suggesting that these models' pre-training may have introduced biases or a tendency to generalize beyond the provided labels. This finding underscores the importance of careful fine-tuning, especially for applications where misclassifications could lead to significant consequences, such as automated decision-making in financial or safety-critical domains.

5.3. Comparative Performance of LLMs and Reasoning Models

In our study, we evaluated eleven LLMs and three reasoning models by prompting them to perform zero-shot classification of consumer complaints into predefined categories. The results of our analysis, based on three key metrics—accuracy, cost, and speed—are summarized in Figure 2.

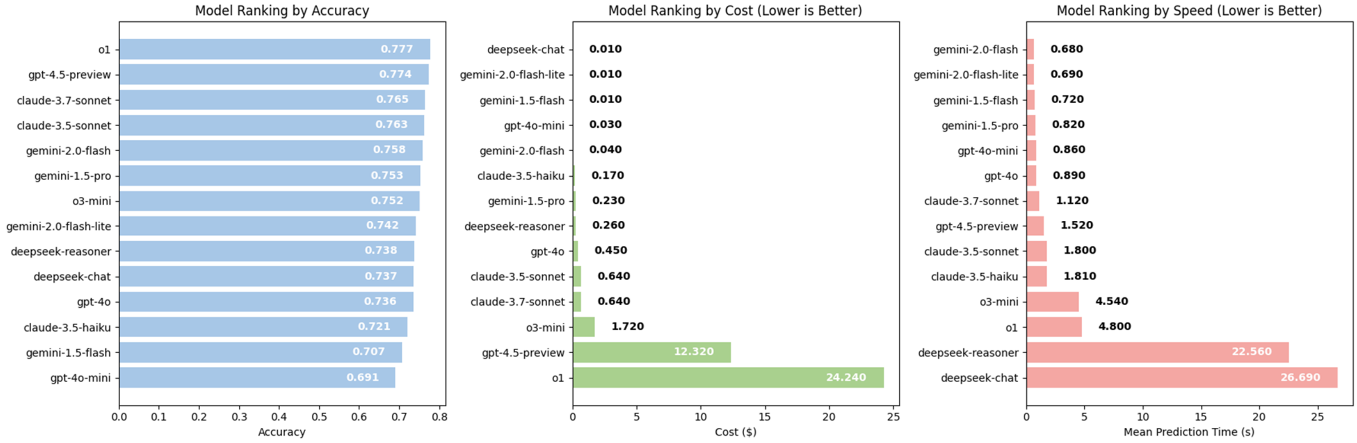

Figure 2. Comparative Classification Performance of LLMs and Reasoning Models Across Three Key Metrics: Accuracy, Cost, and Speed.

Our findings indicate that the reasoning models o1 and o3-mini lead in accuracy, likely due to their extensive chain-of-thought reasoning, which enables deeper processing before generating a response. However, this trend does not extend to DeepSeek-Reasoner, which exhibits nearly identical accuracy to its chat counterpart DeepSeek-Chat. This raises questions about whether DeepSeek-Reasoner truly employs advanced reasoning or merely provides an explanatory breakdown of its conclusions.

Although reasoning models such as o1 and o3-mini dominate in accuracy, Claude-3.5-Sonnet surpasses o1 in precision by 0.29%. Interestingly, this improvement does not carry over to Claude-3.7-Sonnet, its successor, which slightly outperforms Claude-3.5-Sonnet in accuracy by 0.2% but does not maintain the same precision advantage. This suggests that the observed precision gain in Claude-3.5-Sonnet may have been coincidental rather than a direct result of model improvements.

Another key insight is the strong performance of Google's Gemini-2.0-Flash and Flash-Lite models, which outperform GPT-4o in this classification task. Meanwhile, OpenAI's GPT-4.5-preview model demonstrates near-equivalent accuracy to the o1 reasoning model. Given its positioning as OpenAI's latest iteration, it is plausible that GPT-4.5-preview aims to merge the reasoning capabilities of o1 with the general knowledge proficiency of GPT-4o, potentially serving as an all-in-one solution.

Additionally, we observe an almost identical accuracy between GPT-4o (0.736) and DeepSeek-Chat (0.737). This similarity is particularly noteworthy in light of allegations that DeepSeek may have leveraged OpenAI's data for model training [33]. However, despite their comparable accuracy, classification heatmaps reveal distinct behavioral differences. For instance, DeepSeek-Chat did not misclassify complaints into the auto loan category, unlike GPT-4o, suggesting that accuracy alone is insufficient for fully characterizing model behavior. This underscores the need for qualitative analysis alongside traditional performance metrics.

While accuracy is a crucial factor, cost and speed are equally significant in real-world deployment. The o1 model, despite its high accuracy, is among the most expensive and slowest, trailing only the DeepSeek models in response time. This raises an essential question for deployment: should organizations prioritize accuracy at the expense of speed, or seek a balance between the two? Reasoning models, with their extended processing times, may not be ideal for time-sensitive applications. In contrast, Claude models offer a compelling alternative, being approximately 305% faster than o1 while maintaining high accuracy.

From a cost-effectiveness standpoint, Gemini models emerge as strong contenders, offering a balance between affordability and speed, albeit with some trade-offs in accuracy. Model selection ultimately depends on the specific priorities of an application. A graphical representation (Figure 3) illustrates these trade-offs, showing that while the o1 reasoning model remains the most accurate, it is also the most expensive. Gemini-2.0-Flash, in contrast, offers a high-accuracy, cost-effective alternative. Meanwhile, Claude-3.7-Sonnet and Claude-3.5-Sonnet represent a middle ground, balancing accuracy and cost.

In conclusion, selecting the optimal model involves weighing multiple factors beyond accuracy alone. Organizations must carefully consider the trade-offs between accuracy, cost, and speed to align model selection with their specific needs.

5.4. Recommendations for Improving Classification Accuracy

- To enhance classification performance, the following strategies could be employed:
- Fine-tuning on domain-specific financial data to improve model understanding of nuanced complaint categories.
- Incorporating context-aware embeddings to mitigate misclassification in overlapping financial categories.
- Enhancing category definitions to provide clearer distinctions between similar complaint types.

By implementing these improvements, financial complaint classification models can achieve greater accuracy and reliability, ultimately enhancing their applicability in real-world financial analysis and customer service automation.

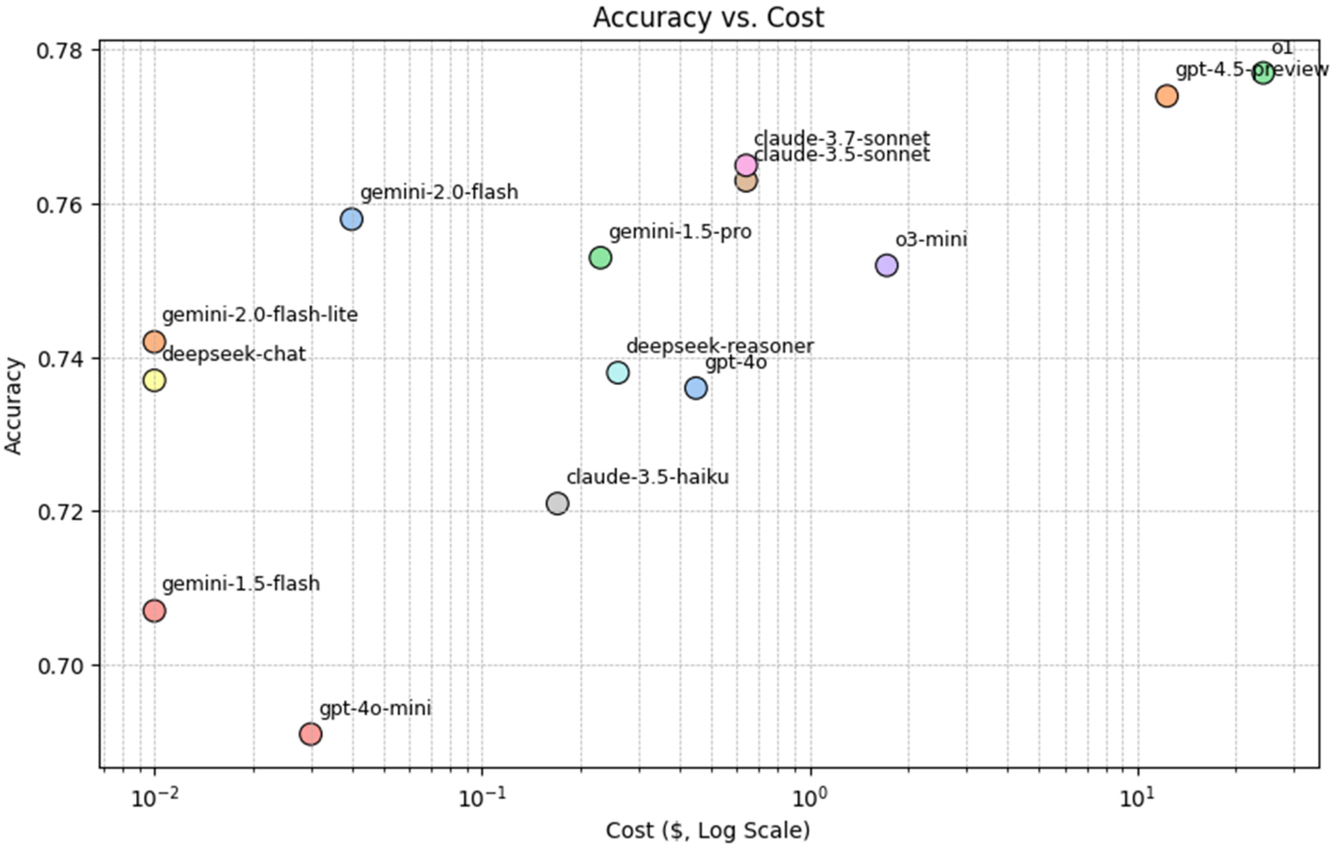

Figure 3. Trade-off Between Accuracy and Cost for LLMs and Reasoning Models.

6. Conclusions

This study represents a groundbreaking step in the application of reasoning models to consumer complaint classification, demonstrating their potential to enhance AI-driven financial text processing. Unlike traditional LLMs, reasoning models such as o1 and o3-mini exhibit advanced inferential capabilities, enabling deeper linguistic and contextual understanding. Our findings reveal that these models achieve superior accuracy in zero-shot classification tasks, marking a significant advancement in automated complaint handling. However, this improvement comes at a cost—reasoning models require longer processing times and higher computational resources, raising practical concerns for real-world deployment.

This study also highlights a fundamental dilemma in model selection: should organizations prioritize accuracy at the expense of cost and speed, or should they seek a balance between these factors? While reasoning models offer state-of-the-art classification performance, their slower response times may not be ideal for time-sensitive applications. Conversely, models such as Claude-3.5-Sonnet and Gemini-2.0-Flash present cost-effective alternatives with competitive accuracy, offering practical trade-offs for businesses. Ultimately, the decision hinges on the specific needs of an organization, whether it be maximizing classification precision, ensuring real-time processing, or optimizing costs. By presenting a comprehensive evaluation of LLMs and reasoning models, this study provides valuable insights into AI-driven financial text analysis, helping organizations navigate the evolving landscape of automated complaint resolution.

Author Contributions: Conceptualization, K.I.R. and N.D.T.; methodology, K.I.R., N.D.T. and D.K.N.; software, K.I.R.; validation, K.I.R. and N.D.T.; formal analysis, K.I.R., N.D.T. and D.K.N.; investigation, K.I.R., N.D.T. and D.K.N.; resources, K.I.R.; data curation, K.I.R.; writing—original draft preparation, K.I.R. and N.D.T.; writing—review and editing, K.I.R., N.D.T. and D.K.N.; visualization, K.I.R.;

supervision, D.K.N. and N.D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data supporting the reported results can be found at [31].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pio, P.G.C.; Sigahi, T.; Rampasso, I.S.; Satolo, E.G.; Serafim, M.P.; Quelhas, O.L.G.; Leal Filho, W.; Anholon, R. Complaint Management: Comparison between Traditional and Digital Banks and the Benefits of Using Management Systems for Improvement. *Int. J. Product. Perform. Manag.* **2024**, *73*, 1050–1070. [CrossRef]
2. Consumer Financial Protection Bureau (CFPB). Consumer Complaint Database. Available online: https://www.consumerfinance.gov/data-research/consumer-complaints/ (accessed on 7 February 2025).
3. Vairetti, C.; Aránguiz, I.; Maldonado, S.; Karmy, J.P.; Leal, A. Analytics-Driven Complaint Prioritisation via Deep Learning and Multicriteria Decision-Making. *Eur. J. Oper. Res.* **2024**, *312*, 1108–1118. [CrossRef]
4. Sharma, S.; Vashisht, M.; Kumar, V. Enhanced Customer Insights: Multimodal NLP Feedback System. In Proceedings of the 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science, Bhopal, India, 24–25 February 2024. [CrossRef]
5. Roy, T.S.; Vasukidevi, G.; Malleswari, T.Y.J.N.; Ushasukhanya, S.; Namratha, N. Automatic Classification of Railway Complaints Using Machine Learning. *E3S Web Conf.* **2024**, *477*, 00085. [CrossRef]
6. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs and NLP Models in Cryptocurrency Sentiment Analysis: A Comparative Classification Study. *Big Data Cogn. Comput.* **2024**, *8*, 63. [CrossRef]
7. Bi, X.; Chen, D.; Chen, G.; Chen, S.; Dai, D.; Deng, C.; Ding, H.; Dong, K.; Du, Q.; Fu, Z.; et al. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv* **2024**, arXiv:2401.02954.
8. Models—OpenAI API. Available online: https://platform.openai.com/docs/models#gpt-4o (accessed on 10 February 2025).
9. Models—Anthropic. Available online: https://docs.anthropic.com/en/docs/about-claude/models (accessed on 10 February 2025).
10. Gemini API Gemini Models. Gemini API. *Google AI for Developers.* Available online: https://ai.google.dev/gemini-api/docs/models/gemini (accessed on 2 March 2025).
11. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Fake News Detection and Classification: A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models. *Future Internet* **2025**, *17*, 28. [CrossRef]
12. OpenAI Reasoning Models—OpenAI API. Available online: https://platform.openai.com/docs/guides/reasoning (accessed on 2 March 2025).
13. Sakas, D.P.; Reklitis, D.P.; Terzi, M.C.; Glaveli, N. Growth of Digital Brand Name through Customer Satisfaction with Big Data Analytics in the Hospitality Sector after the COVID-19 Crisis. *Int. J. Inf. Manag. Data Insights* **2023**, *3*, 100190. [CrossRef]
14. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Leveraging Large Language Models in Tourism: A Comparative Study of the Latest GPT Omni Models and BERT NLP for Customer Review Classification and Sentiment Analysis. *Information* **2024**, *15*, 792. [CrossRef]
15. Kumar, P.; Singh, A.; Saha, S. Navigating the Indian Code-Mixed Terrain: Multitasking Analysis of Complaints, Sentiment, Emotion, and Severity. 2024. Available online: https://ssrn.com/abstract=4827145 (accessed on 10 February 2025).
16. Seok, J.; Kim, C.; Kim, S.; Kim, Y.-M. Deep-Learning-Based Customer Complaints Monitoring System Using Online Review. 2024. Available online: https://ssrn.com/abstract=4795530 (accessed on 10 February 2025).
17. Khadija, M.A.; Nurharjadmo, W. Enhancing Indonesian Customer Complaint Analysis: LDA Topic Modelling with BERT Embeddings. *SINERGI* **2024**, *28*, 153–162. [CrossRef]
18. Juipa, A.; Guzman, L.; Diaz, E. Sentiment Analysis-Based Chatbot System to Enhance Customer Satisfaction in Technical Support Complaints Service for Telecommunications Companies. In Proceedings of the International Conference on Smart Business Technologies (ICSBT), Dijon, France, 8–10 July 2024.
19. Das, S.; Singh, A.; Saha, S.; Maurya, A. Negative Review or Complaint? Exploring Interpretability in Financial Complaints. *IEEE Trans. Comput. Soc. Syst.* **2024**, *11*, 3606–3615. [CrossRef]
20. Jia, S.; Shan, G.; Chi, O.H. Leveraging Generative AI for Customer Complaint Resolution: A Comparative Analysis with Human Responses. In Proceedings of the AMCIS 2024, Salt Lake City, UT, USA, 15–17 August 2024.
21. Correa, N.; Correa, A.; Zadrozny, W. Generative AI for Consumer Communications: Classification, Summarization, Response Generation. In Proceedings of the IEEE Andescon, Cusco, Peru, 11–13 September 2024. [CrossRef]

22. Jondhale, R.; Patil, S.; Shinde, A.; Ajalkar, D.; Biradar, S. Predicting Consumer Complaint Disputes in Finance Using Machine Learning (AIOPS). In Proceedings of the 2nd IEEE International Conference on Advances in Information Technology, ICAIT, Chikkamagaluru, India, 24–27 July 2024. [CrossRef]
23. Ismailbekovich, B.D. Implementing chatbots for consumer complaint response. *Proc. Int. Conf. Mod. Sci. Sci. Stud.* **2024**, *3*, 440–445. Available online: https://econferenceseries.com/index.php/icmsss/article/view/3992 (accessed on 9 February 2025).
24. Song, W.; Rong, W.; Tang, Y. Quantifying Risk of Service Failure in Customer Complaints: A Textual Analysis-Based Approach. *Adv. Eng. Inform.* **2024**, *60*, 102377. [CrossRef]
25. Wang, R.; Wang, H.; Li, S. Predicting the Determinants of Consumer Complaint Behavior in E-Commerce Live-Streaming: A Two-Staged SEM-ANN Approach. *IEEE Trans. Eng. Manag.* **2025**. *early access*. [CrossRef]
26. Zhou, X.; Cao, G.; Peng, B.; Xu, X.; Yu, F.; Xu, Z.; Yan, Y.; Du, H. Citizen Environmental Complaint Reporting and Air Quality Improvement: A Panel Regression Analysis in China. *J. Clean. Prod.* **2024**, *434*, 140319. [CrossRef]
27. Lee, C.H.; Zhao, X. Data Collection, Data Mining and Transfer of Learning Based on Customer Temperament-Centered Complaint Handling System and One-of-a-Kind Complaint Handling Dataset. *Adv. Eng. Inform.* **2024**, *60*, 102520. [CrossRef]
28. Tiwari, S. Consumer Complaints Dataset for NLP. Available online: https://www.kaggle.com/datasets/shashwatwork/consume-complaints-dataset-fo-nlp (accessed on 7 February 2025).
29. Zhang, K.; Zhou, F.; Wu, L.; Xie, N.; He, Z. Semantic Understanding and Prompt Engineering for Large-Scale Traffic Data Imputation. *Inf. Fusion.* **2024**, *102*, 102038. [CrossRef]
30. Anthropic PBC. Automatically Generate First Draft Prompt Templates. Available online: https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/prompt-generator (accessed on 30 November 2024).
31. Roumeliotis, K. GitHub—Applied-AI-Research-Lab/DeepSeek-LLM-and-GPT-Fall-Behind-Claude-Leads-in-Zero-Shot-Consumer-Complaints-Classification. Available online: https://github.com/Applied-AI-Research-Lab/DeepSeek-LLM-and-GPT-Fall-Behind-Claude-Leads-in-Zero-Shot-Consumer-Complaints-Classification (accessed on 7 February 2025).
32. Zhao, S.; Guo, Y.; Sheng, Q.; Shyr, Y. Advanced Heat Map and Clustering Analysis Using Heatmap3. *Biomed. Res. Int.* **2014**, *2014*, 986048. [CrossRef] [PubMed]
33. Cade Metz OpenAI Says DeepSeek May Have Improperly Harvested Its Data. *The New York Times*, 29 January 2025. Available online: https://www.nytimes.com/2025/01/29/technology/openai-deepseek-data-harvest.html (accessed on 9 February 2025).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.